# 哈工大语言技术研究中心 CWMT2009 机器翻译评测技术报告

张春越 蒋宏飞 刘水 刘宇鹏 王博 许冲 赵铁军
哈尔滨工业大学 语言技术研究中心 哈尔滨 150001
E-mail: {cyzhang,hfjiang,liushui,ypliu,bowing,xuchong,tjzhao}@mtlab.hit.edu.cn

**摘 要**: 本文介绍了哈尔滨工业大学语言技术研究中心参加 CWMT09 机器翻译评测的情况。我们参加了汉英新闻单一系统、汉英新闻系统融合、英汉新闻机器翻译和英汉科技机器翻译 4 个项目。在评测中，我们使用了 HiTree 和 Water 两个统计机器翻译系统、两个词一级融合系统（WordComb1 和 WordComb2）和一个句子一级融合系统 wMBR。本文对各个系统进行了简要的介绍，并给出了各个系统参加评测时的数据配置情况和最终的评测结果。

**关键字**: 统计机器翻译、系统融合、句法分析、同步树替换文法、混合同步文法

# Technical Report of HIT_LTRC for CWMT 2009 Evaluation

Chunyue Zhang,Hongfei Jiang,Shui Liu,Yupeng Liu,Bo Wang,Chong Xu,Tiejun Zhao
Language Technology and Research Center
Harbin Institute of Technology
Harbin, China, 150001
E-mail: {cyzhang,hfjiang,liushui,ypliu,bowing,xuchong,tjzhao}@mtlab.hit.edu.cn

**Abstract:** *This paper describes the systems submitted to the China Workshop on Machine Translation 2009 by language technology and research center in HIT. We participated in translation tracks(ZH-EN_NEWS-SINGL, EN-ZH-NEWS-TRANS, EN-ZH-SCIE-TRANS) and system combination track(ZH-EN-NEWS-COMBI). In the evaluation campaign, we used 2 statistical machine translation systems (HiTree and Water), two word-level system combiners and one sentence-level combiner (wMBR). This paper briefly describes each system and presents detailed experiment setup and test scores.*

**Keywords:** *statistical machine translation, system combination of machine translation, parsing, Synchronous tree substitution grammar, Synthetic Synchronous Grammar*

## 1 引言

哈尔滨工业大学语言技术研究中心隶属于哈尔滨工业大学计算机科学与技术学院。本中心一直致力于包括机器翻译在内的自然语言处理及相关人工智能理论与技术的研究及相关系统的研制和开发。我们于 1989 年开发完成国内第一个机器翻译系统，之后我们陆续开发了多个基于规则与基于实例的机器翻译系统。近年来，我们在统计机器翻译领域开展了新的工作，并连续参加了 SSMT2007、CWMT2008 和 CWMT2009 三年的机器翻译评测。

此次评测，我们参加了除汉蒙日常用语机器翻译评测外 4 个项目的评测，包括汉英新闻领域单一系统评测，英汉新闻领域机器翻译评测，英汉科技领域机器翻译评测及汉英新闻领域的系统融合评测。下面对我们参加评测的系统和系统使用的数据以及相关实验进行简要的说明。

## 2 参评系统描述

在这次评测中，我们使用了以下系统：

1. HiTree，基于混合同步文法的统计机器翻译系统
2. Water，基于短语模型的统计机器翻译系统
3. WordComb1，词一级融合系统
4. WordComb2，词一级融合系统
5. wMBR，句子一级融合系统

### 2.1 HiTree

HiTree 是一种基于混合同步文法(Synthetic Synchronous Grammar, SSG)的统计句法系统。在本次评测中，HiTree 融合了形式化的上下文无关文法(Synchronous Context-Free Grammar, SCFG)和基于语言学句法信息的同步树序列替换文法(Synchronous Tree Sequence Substitution Grammar, STSSG)。图 1 给出了一个带有词对齐信息的句法树树对的样例，图 2 列出了一些从图 1 所示句法树树对中可以抽取出的混合文法规则。从图 2 中可以看出，在 HiTree 目前的实现中，可以综合利用短语规则，纯形式化的 SCFG 规则，以及 STSSG 规则。关于 HiTree 的更多细节，请参看参考文献[Jiang et al.，2009]。

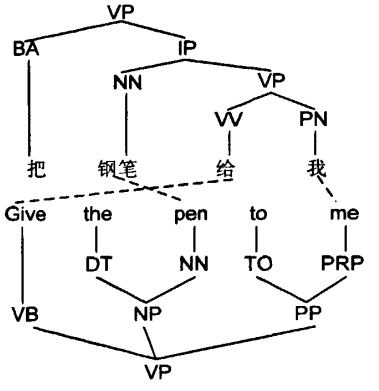


**图 1. 带有词对齐信息的句法树树对样例**

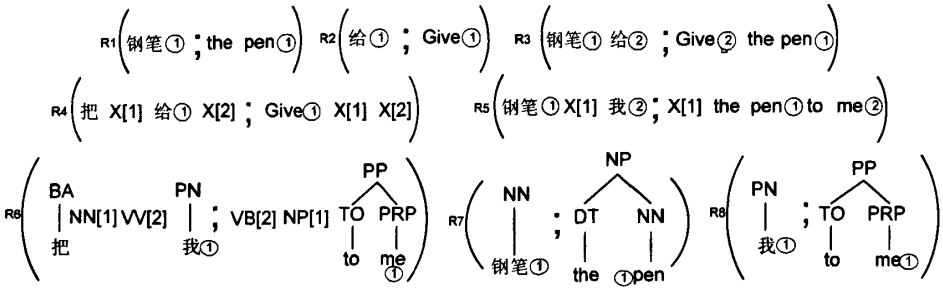


**图 2. 从图 1 中可以抽取出的 HiTree 规则。特别地，R1-R3 是短语翻译规则，R4-R5 是纯形式化的 SCFG 规则，R6-R8 是 STSSG 规则.**

因为 HiTree 的训练句子需要包含句法信息，而句法分析的处理时间非常耗时，所以在本次评测中，本系统仅使用了训练语料中 2,686,888 个较短的句对。以我们单位实现的另一短语系统 Water 作为基线系统，此系统在内部测试中性能类似于 Moses,此次评测采用了 3,745,542 个训练句对，得到的本次评测结果对比表 1 所示:

表 1 HiTree 和 Water 的对比结果

| 项目 | 参评系统 | BLEU-SBP |
|---|---|---|
| 汉英新闻领域单一系统 | HiTree | 0.2088 |
| | Water | 0.1901 |

从表 1 的对比结果可以看出，HiTree 在使用了 70%的训练语料的情况下，仍然可以取得较短语系统明显的优势，证明了句法系统的有效性。

另外，WoodPecker 的评测结果显示，HiTree 系统在汉英新闻单系统项目中总体排名第四，在 Sentences 类型检测点上的排名第一。这些结果显示，HiTree 系统输出译文的句法合法性更合理一些，特别是在句子级上。

表 2

| Score | systran | ia | i2r | hit |
|---|---|---|---|---|
| GeneralScore | 0.2981 | 0.2933 | 0.2862 | 0.2861 |
| Words | 0.5066 | 0.4883 | 0.4729 | 0.4696 |
| Phrases | 0.2786 | 0.2751 | 0.2686 | 0.2689 |
| Sentences | 0.2843 | 0.1633 | 0.2653 | **0.3163** |

## 2.2 Water

Water 系统是类似于 Moses[Koehn et al., 2007]的一个基于短语的统计机器翻译系统。该系统将信道模型应用到机器翻译模型中，应用 Viterbi 算法进行解码，在 Viterbi 解码过程中，剪枝策略是影响系统性能和运行效率的主要原因之一，相比于 Moses ，本系统实现将多种剪枝策略相融合成一种综合的剪枝策略。以确保该算法的解码质量。

**Algorithm 1** beam search with histogram pruning and recombine

```
procedure ADD2STACK(h, s)
    if empty(s) then
        insert(h, s)
        return
    else if f(h) < max(s)/HW then
                                    ▷ beam − search
        return
    else if f(h) < min(s) and ! eq(h, s) then
            ▷ histogram pruning and recombine
        if size(s) < HW then
            insert(h, s)
        end if
        return
    else if f(h) > min(s) and ! eq(h, s) then
            ▷ histogram pruning and recombine
        if size(s) = HW then
            eraseMin(s)
        end if
        insert(h, s)
        return
    else if be(h, s) then
        eraseEqual(h, s)                ▷ recombine
        insert(h, s)
        return
    end if
end procedure
```

图 3 Water 系统剪枝算法

单纯基于重合并的剪枝算法可以在一定程度上减少解码算法在寻优过程中的代价，但是并不能将解码的时间复杂度降低至多项式时间，为了将解码代价降低至多项式时间，在普通

解码模式下通常采用束搜索(beam-search)剪枝技术和柱状图剪枝技术(histogram pruning)以达到这个目的。与重合并剪枝技术不同，这两种剪枝技术都存在将潜在的最优路径丢弃的风险。

柱状图剪枝技术，只保存当前最优的 N 条路径(N 为柱状图剪枝的阀值)，丢弃在这个范围以外的搜索路径，由此，可以证明该剪枝策略将解码算法的代价控制在多项式间内。

束搜索剪枝是一种在许多寻优技术中广泛应用的剪枝技术，该剪枝技术将目前路径的最高的得分除以某个大于 1 的阀值，所有在此阀值以下的路径都将被丢弃。由于当前最高的得分未必是最终的最优得分，束搜索剪枝是一种与搜索路径顺序相关的剪枝技术。

考虑到以上剪枝策略在解码过程中可能同时丢弃同一条翻译路径，并在一定程度上相互影响，为了在搜索过程中尽量避免这种情况，同时实现以上算法的逻辑，将以上 3 种策略融合成一种混合策略，最终的剪枝算法的伪代码如图 3 所示。

其中，图 3 所示的算法中：

- s 为存储当前翻译路径 h 的数据结构，HW 为柱状图剪枝的阀值，BW 为束剪枝的阀值。
- 函数 empty(s)当 s 中没有记录任何翻译路径时返回 true，否则返回 false。
- 函数 insert(h,s)将翻译路径 h 存储到 s 中。
- 函数 f(h)返回路径 h 的翻译得分。
- 函数 max(s)返回在 s 中得分最高的翻译路径 h。
- 函数 min(s)返回在 s 中翻译路径的最低得分。
- 函数 eraseMin(s)将删除 s 中得分最低的翻译路径。
- 函数 eraseEqual(h,s)函数 size(s)返回 s 存储的翻译路径数。
- 函数 eq(h,s)当 s 中存在符合与 h 进行重合并条件的翻译路径时返回 true，否则返回 false。
- 函数 be(h,s)当 s 中存在符合与~$h$~做重合并条件的翻译路径并且该路径的得分低于 h 时，返回 true，否则返回 false。

此外，cube-pruning 是广泛应用在完全句法分析和机器翻译中的一种高效解码方法，为了提高系统的速度，本系统实现了该算法，使系统速度得到大幅度的提升。关于更多 Water 的实现细节，请见参考文献[刘水 等.2009]。

## 2.3 WordComb1

本系统的思想来自于[Sim et al.,2007;Rosti et al.,2007]，并在基础上增加了增量策略。假如的方法如下：

1. 一个系统的输出（通常是带有最好词序或是性能最好系统的 top1）选做骨架翻译。这里的所有词作为和其他机器翻译结果对齐的锚点。一个初始的混淆网络被创建。每个“词袋中含有一个词”。
2. 把系统的输出结果和混淆网络一步一步的对齐。这个对齐是用 TER 来进行对齐的，它允许系统输出关于混淆网络进行重排序。这里的词串和“词袋串”匹配定义为词串和“词袋串”中的任意路径进行匹配。
3. 如果词需要插入到混淆网络中，一个新的“词袋”被创建含有两个弧：一个是它的标签为插入词，另一个是带有特殊标签“NULL”。如果一个“词袋”获得删除操作，那么“NULL”弧被插入到“词袋”中。匹配和替代操作的词都插入到“词袋”中。
4. 弧的代价是全部系统对该词代价的后验概率和。来自 N-best 列表中的词获得 1/（1+m）的值，这里 m 是假设在系统中的排名。
5. 最后混淆网络接着用 5 元语言模型来进行打分。在线性组合中的权重通过开发集来决定。

WSD 的加入，为了把基于 WSD 的方法加入到增量的 TER 算法中，我们对 TER 算法进行修

改，把原来的替代操作的值改成由 WordNET 计算的值。
解码的公式如下：

$$\log p(E_{j,n} \mid F_j) = \sum_{i=1}^{N_j-1} \log(\sum_{l=1}^{N_l} \lambda_l p(w \mid l,i)) + \nu L(E_{j,n}) + \mu N_{null}(E_{j,n}) + \xi N_{words}(E_{j,n})$$

更多关于 WordComb1 的实现细节，请见参考文献[刘宇鹏 等. 2009]。

## 2.4 WordComb2

本系统是在原来 T E R 对齐的基础上加入了启发式规则和增量策略，加入启发式规则的方法如下：

1. 产生两个方向的词对齐结果。是直接由 GIZA 或是 IHMM 生成的词对齐结果。这个可以获得通过切换源语言和目标语言的参数来获得。由于 TER 算法的贪心搜索的特性，通过转换源语言和目标语言的参数,双向的 TER 的对齐结果不必要时完全相同的。当两个词对齐已经准备好了后了，我们从两个词对齐的交集开始，接着在骨架翻译和假设加入新的链，当且仅当新链的两个词没有对齐，同时这个链存在于两个词对齐的交集中。如果有两个链共享一个骨架翻译或是假设翻译词，也满足于限制。我们选择最高相似分数，相似度分数使用最大公共子串来计算，计算公式如下：

$$S(e_i, e_j) = \frac{2 \times len(MCS(e_i, e_j))}{len(e_i) + len(e_j)}$$

基于增量算法的 TER 对齐的启发式算法有所不同，因为现在的骨架翻译是一个混淆网络，那么就需要在生成的 TER 对齐是假设翻译的一个词和骨架翻译的一袋子词。而在扩展的过程中，出现有两个链共享一个骨架翻译或是假设翻译词，采用一个词和一袋子词中 MCS 中最高分作为该得分。当然其他的情况与（2）中计算方法相同。

更多关于 WordComb2 系统的细节，请参考文献[刘宇鹏 et al.2009]。

## 2.5 wMBR

本系统的思想来源于[Sim，2007]。其基本内容是在全部N-best译文中找到与全体译文相比具有最小贝叶斯风险(Minmum Bayes Risk)的译文作为最佳输出译文。设$E_{mbr}$为最佳译文，则：

$$E_{mbr} = \arg\min_{E'} \sum_{E} P(E \mid F) L(E, E')$$

其中 $L(E,E')$ 用来度量两个译文间的期望风险。在本系统中，我们使用译文间的 BLEU 分数的倒数作为风险的度量。在计算 BLEU 分数时，$E'$作为待测译文，$E$ 作为参考译文。

$$L(E, E') = 1 / BLEU(E, E')$$

$P(E|F)$ 为译文 $E$ 的后验概率：

$$P(E \mid F) = \frac{P(E,F)}{\sum_{E'} P(E',F)}$$

这里我们使用译文的系统分数(system score) 来度量联合概率 P(E,F)。

在具体实现中，我们为每个系统译文根据其所在系统的翻译质量进行加权。系统的翻译

质量用系统在开发集合上的BLEU分数来度量，设$W_E$ 为译文$E$的权重：

$$W_E = BLEU_i(DevData)$$

其中 $i$ 为 $E$ 所在系统的序号。加权后选取最佳译文的公式为：

$$E_{mbr} = \frac{1}{W_{E'}} \arg\min \sum_{E} W_i P(E|F) L(E,E')$$

# 3 实验结果

## 3.1 数据使用

此次评测我们参加了汉英新闻领域单一系统评测，英汉新闻领域机器翻译评测，汉英科技领域机器翻译评测及汉英新闻领域的系统融合评测。使用的语料均为本次评测允许使用的语料。

具体使用的数据集如表 3 所示：

表 3 主办方提供的汉英训练数据

| 资源描述 | 双语句对数量 |
|---|---|
| CLDC-LAC-2003-004 | 252327 |
| CLDC-LAC-2003-006 | 200082 |
| 厦门大学英汉电影字幕平行语料 | 176148 |
| 哈工大信息检索组汉英句子级对齐语料库 | 100000 |
| 哈工大机器智能与翻译研究室英汉对齐语料 | 52227 |
| 点通汉英平行语料库（部分） | 1000004 |
| 计算所 Web 汉英平行语料库 | 1076313 |
| 万方汉英中文科技论文摘要语料库 | 320984 |
| 中信所英汉科技文献句子级对齐语料库 | 611527 |
| 总计 | 3789612 |

## 3.2 数据处理方法及工具

中文分词的处理工程使用了 Stanford 的中文分词工具[Tseng et al., 2005.]进行汉语分词。

为了获得 HiTree 需要的双语句法分析结果，我们使用了 Stanford 的中英文句法分析工具进行句法分析[Klein et al., 2003; Levy et al. 2003]。

词语对齐工具使用 GIZA++[Och et al., 2003]。双语语料完成双向词对齐后，采用启发式规则 grow-diag-final 合并两个方向的词对齐结果，作为最终的双语词对齐结果。

语言模型训练工具使用 SRILM[Stolcke, 2002]。

对于翻译结果，我们仅进行了未登录词删除处理。

## 3.3 各系统具体情况

在参加本次评测各个项目时，我们分别使用了如下系统：

1）汉英新闻领域单一系统：HiTree 和 Water

2）英汉新闻领域机器翻译：Water 和 wMBR

3) 英汉科技领域机器翻译：Water

4) 汉英新闻领域系统融合：WordComb1、WordComb2 和 wMBR

在 1） 2） 3）三个项目中，我们用以下规则对表 1 中的数据进行过滤，最终使用的训练语料为 3745542 句。

1）双语的句子长度均不超过 80 个词

2）双语的句子长度比不超过 7

在汉英新闻领域单一系统项目中，Water 使用了全部 3,745,542 句训练语料，解码方式为普通解码。HiTree 由于受句法分析的限制，只使用了表 3 中除去万方和中信所外的新闻语料部分，且双语句子长度不超过 50 个词，双语句子长度比不超过 7，共 2,686,888 句。Water 和 HiTree 使用的语言模型相同，都使用了评测组织方提供的训练语料中的英文部分以及允许使用的非评测组织方提供的训练数据——路透社语料库第一卷，用 SRILM 工具训练了一个 5 元语言模型。开发集使用的是从 CWMT2008 汉英新闻测试集（共 1006 句）中随机选取的 500 句。

在英汉新闻领域机器翻译项目中，Water 使用了全部 3,745,542 句训练语料，中文语言模型使用了评测组织方提供的训练语料中的中文部分以及允许使用的非评测组织方提供的训练数据——搜狗全网新闻语料库，用 SRILM 工具构建了 5 元语言模型。按照解码方式分别为普通解码和 cube-pruning 解码给出了 2 组翻译结果，开发集使用的是从 2005-863-001 评测语料里随机选取的 513 句；wMBR 则融合了这 2 组翻译结果（各 10-best）。其中，wMBR 的融合结果作为主系统提交，Water 用不同的两种解码方式得到的翻译结果分别作为 Contrast-systemb 和 Contrast-systemc 系统。

在英汉科技领域机器翻译项目中，我们只有 Water 一个系统参加评测。Water 使用了全部 3,745,542 句训练语料。语言模型的使用同英汉新闻机器翻译项目。按照解码方式分别为普通解码和 cube-pruning 解码给出了 2 组翻译结果。开发集使用的是从 CWMT2008 英汉科技机器翻译测试集（1002 句）中随机选取的 500 句。其中，普通解码方式作为 Primary 系统，cube-pruning 解码方式作为 Contrast 系统。

在汉英新闻系统融合项目中，WordComb1 和 WordComb2 使用了在汉英新闻领域单一系统项目中的语言模型，开发集为会议评测方提供的 SSMT07 系统译文的 10best 结果。wMBR 按照使用的开发集的不同给出了两组结果：第一组的开发集使用的是评测方发放的 SSMT07 系统译文的 10best 结果，BLEU-SBP 得分前 9 位的系统作为输入系统。测试语料使用会方提供的 CWMT09 系统译文的 10best 结果，BLEU-SBP 得分前 9 位的系统作为输入系统；第二组的开发集使用评测放发放的 SSMT07 系统译文的 10best 结果，BLEU-SBP 得分前 12 位的系统作为输入系统。测试语料使用会方提供的 CWMT09 系统译文的 10best 结果，BLEU-SBP 得分前 12 位的系统作为输入系统。最终，WordComb1 作为 Primary-systema 系统，WordComb2 作为 Contrast-systemb 系统， wMBR 的第一组结果作为 Contrast-systemc 系统，wMBR 的第二组结果作为 Contrast-systemd 系统。

### 3.4 评测成绩

表 4 给出了我们参加 CWMT2009 评测的正式成绩。我们分别以 HiTree（汉英新闻领域单一系统）、WordCombo1（汉英新闻领域系统融合）、wMBR（英汉新闻领域机器翻译）和 Water（英汉科技领域机器翻译）作为主系统。

表 4 CWMT2009 评测成绩

| 项目 | 参评系统 | 类别 | BLEU-SBP |
|---|---|---|---|
| 汉英新闻领域单一系统 | HiTree | Primary-systema | 0.2088 |
| | Water | Contrast-systemb | 0.1901 |
| 汉英新闻领域系统融合 | WordComb1 | Primary-systema | 0.2422 |
| | WordComb2 | Contrast-systemb | 0.2448 |
| | wMBR | Contrast-systemc | 0.2475 |
| | wMBR | Contrast-systemd | 0.2473 |
| 英汉新闻领域机器翻 | wMBR | Primary-systema | 0.3179 |

| | | | |
|---|---|---|---|
| 译 | water | Contrast-systemb | 0.3217 |
| | water | Contrast-systemc | 0.3157 |
| 英汉科技领域机器翻译 | water | Primary-systema | 0.4497 |
| | water | Contrast-systemb | 0.4418 |

从评测成绩来看，我们主要可以得到以下两个结论：

1）基于混合同步文法的 HiTree 系统在利用部分语料情况下仍然在汉英新闻单一系统项目上取得了不错的成绩,证明了句法模型的有效性。

2）基于句子级的融合系统取得了较好的成绩，但令人感到意外的是基于词一级的融合系统并没有取得稳定的优势性能，可能的原因是 N-Best 译文噪音较大。

## 4 总结

本文主要介绍了哈尔滨工业大学语言技术研究中心参加第五届统计机器翻译评测活动的统计机器翻译系统和融合系统的实现情况、实验配置和测试结果。我们参加了汉英新闻领域单一系统评测，英汉新闻领域机器翻译评测，汉英科技领域机器翻译评测及汉英新闻领域的系统融合评测，在系统融合评测中取得了较好的成绩。

## 参考文献


Hongfei Jiang, Muyun Yang, Tiejun Zhao, Sheng Li and Bo Wang. A Statistical Machine Translation Model Based on a Synthetic Synchronous Grammar. ACL-2009, short paper.

Philipp Koehn, Hieu Hoang. Moses: Open Source Toolkit for Statistical Machine Translation. Proceedings of the ACL 2007 Demo and Poster Sessions, pages 177–180,Prague, June.

Antti-Veikko I. Rosti etc. 2002. Combining Outputs from Multiple Machine Translation Systems, NAACL 2002.

K.C.Sim etc.,2007. Consensus network decoding for statistical machine translation system combination, ICASSP 2007.

Franz Josef Och, Hermann Ney. "A Systematic Comparison of Various Statistical Alignment Models", Computational Linguistics, volume 29, number 1, pp. 19-51 March 2003

A. Stolcke.2002. SRILM-An extensible language modeling toolkit.In Porceedings of the international Conference on Spoken Language Processing,pp.901-904

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky and Christopher Manning. "A Conditional Random Field Word Segmenter." In Fourth SIGHAN Workshop on Chinese Language Processing. 2005.

Dan Klein and Christopher D. Manning. 2003. Fast Exact Inference with a Factored Model for Natural Language Parsing. In Advances in Neural Information Processing Systems 15 (NIPS 2002), Cambridge, MA: MIT Press, pp. 3-10.

Roger Levy and Christopher D. Manning. 2003. Is it harder to parse Chinese, or the Chinese Treebank?. ACL 2003.

刘水，李生，赵铁军，张春越，王博. 基于剪枝策略的短语机器翻译算法研究.第五届全国机器翻译研讨会论文集.2009

刘宇鹏，赵铁军，李生. 混淆网络中对齐策略的研究.第五届全国机器研讨会论文集.2009